IJHSR
ijhighschoolresearch.org
■ RESEARCH ARTICLE

Prediction of Lung Cancer Survival using Machine Learning Algorithms

Patrick Junhee Kim, Andrew Seungwoo Youn

Asia Pacific International School (APIS), 57 Wolgye-ro 45ga-gil, Nowon-gu, Seoul, 01874, South Korea; hipatrick18@gmail.com
Mentor: Dr. Zuhawn Sung

ABSTRACT: Recently, machine learning (ML), a subfield of artificial intelligence (AI), thrives in the healthcare system. It is widely accepted by healthcare practitioners that ML can revolutionize the decision-making procedure on diagnosis and prognosis of disease. ML reveals its importance by detecting key features from complex datasets based on routine clinical and laboratory evaluations. Through this process, predictive accuracy is computed, and reliable results are provided with the estimations of generalization errors. Significantly, such technical advantages of ML become further significant in cancer research. A more accurate prediction can provide curative treatments and prevent death by managing clinical procedures. In this study, we demonstrate the methods to predict the survival rate of advanced lung cancer patients by utilizing supervised machine learning methods. We identified the key features strongly associated with cancer survival from a publicly accessible dataset from the North Central Cancer Treatment Group (NCCTG). We hope this study follows up on the potential of ML technical applications used to prevent lung cancer death.

KEYWORDS: Biomedical and Health Sciences; Lung Cancer; Machine Learning; Non-small Cell lung Cancer; Exploratory Data Analysis; Random Forest and KNN models.

■ Introduction

Over the past few decades, the life expectancy of humans has increased alongside the advancement in medical technology. Recently, this trend has been further raised by coupling with Artificial Intelligence (AI) and Machine Learning (ML) techniques.[1] Their applications are not limited to clinical practice, traditional medicine, and biomedical research of various diseases.[2] Many different ML tools have been developed to help healthcare organizations: achieve automated routine tasks, improve decision-making, diagnose accurately, enhance the patient experience, accelerate innovations, reduce operating costs, reduce risks, increase staff satisfaction, and so on.[3] More importantly, Al and its subfield have made eminent progress in cancer research within oncology. ML tools compute a large volume of the dataset based on routine clinical and laboratory measures to find a reliable predictive outcome strongly associated with the key features that lead to impressive performance like highly accurate diagnoses for preventing cancer deaths. ML tools can improve the accuracy of cancer susceptibility, recurrence, and survival prediction. ML-powered decision-making leads to advanced prognosis and prediction in managing clinical treatments for cancer diseases.[4] In turn, the 5-year cancer survival rate increases. A recent review[5] addressed that the prediction accuracy of cancer development has been increased by 15-20% with ML techniques.

Cancer is a disease in which some of the body's cells grow uncontrollably and spread to other body parts. On average, 1 in 2 men and 1 in 3 women are diagnosed with cancer.[6] Early detection of localized stages of cancer provides additional opportunities to deliver curative treatment and prevent deaths. Multiple ML algorithms are implemented to draw insights from large medical datasets of cancer patients and teach themselves how to analyze and interpret data to help clinicians with evaluation, decision, prevention, and treatment.[1] Although these outcomes are promised, utilizing ML algorithms and integrating them into clinical practice confronts technological challenges in model development.[4] In general, such a challenge arises from a lack of data validation due to patient privacy and medical sensitivity of data features. Time limitations of clinical measures and biased parameters from the medical registry can also degrade the dataset's quality.[4]

Accuracy is a measure related to the total number of correct predictions. To realize reliable practical achievements regarding the predicting performance of a model, various models need to be correlatively compared to determine the accuracy level. In this study, we attempt to predict lung cancer survival rates using supervised machine learning algorithms based on the medical dataset from the North Central Cancer Treatment Group (NCCTG), which is publicly available.[7] This dataset is the record of the patients who were diagnosed with advanced lung cancer. From this work, we would like to demonstrate that machine learning is an enabling tool to produce effective and accurate predictions in cancer and is a future clinical method to provide manageable clinical treatments for cancerous conditions to prevent deaths.

■ Method

Dataset:

Worldwide, lung cancer is the second most commonly diagnosed cancer, contributing 12.2% of the total number of new cases diagnosed in 2020.[8] It is the most common cancer in men, contributing 15.4% of the overall cancer diagnoses. Lung cancer is grouped into two main types; small cell and

DOI: 10.36838/v5i5.3

cell (including adenocarcinoma and squamous cell carcinoma) lung cancer.[9] Non-small cell lung cancer (NSCLC) is the most common type in the United States, accounting for 82% of all lung cancer diagnoses.[9] The 5-year survival rates of this cancer for men and women are 18% and 25%, respectively. The survival rate for non-small cell lung cancer (NSCLS) is 26%, compared to 7% for small cell lung cancer.[10] Like other cancers, lung cancer can be cured in several clinical ways, depending on the type or stage and how far it has spread; for instance, surgery, chemotherapy, radiation therapy, and targeted therapy.[10] More details on the oncological information of lung cancer can be found in the references.[8-10]

The North Central Cancer Treatment Group (NCCTG) dataset records the lung cancer survivors' assessments for the performance status measured by the physician and by the patients. It evaluates how patients' self-assessment can provide prognostic information complementary to the physician's assessment. The dataset contains 228 patients, including 63 patients that are right-censored (patients who left the study before their death). It can be downloaded via Datxplore demos.[11] Figure 1 shows a part of the NCCTG data set; "Inst" is the institute ID for the patient clinics, "time" is the period of the observation, "status" is the indication of the patient's status (1 = censored, 2 = death), "ph.ecog" is the performance status assessed by the physician based on the eastern cooperative oncology group (ECOG) with a scale of 0 - 5 (0 = fully active, 5 = death), "ph.karno" is the Karnofsky performance status, assessed by the physician, on a scale from 0 (death) to 100 (completely healthy), "pat.karno" is the Karnofsky performance status, access by the patient, on a scale from 0 (death) to 100 (completely healthy), "meal.cal" is the average meal calories, consumed by the patient, and "wt.loss" is the amount of weight loss of the patient in last six months.

	Unnamed: 0	inst	time	status	age	sex	ph.ecog	ph.karno	pat.karno	meal.cal	wt.loss
0	1	3.0	306	2	74	1	1.0	90.0	100.0	1175.0	NaN
1	2	3.0	455	2	68	1	0.0	90.0	90.0	1225.0	15.0
2	3	3.0	1010	1	56	1	0.0	90.0	90.0	NaN	15.0
3	4	5.0	210	2	57	1	1.0	90.0	60.0	1150.0	11.0
4	5	1.0	883	2	60	1	0.0	100.0	90.0	NaN	0.0

Figure 1: NCCTG Data set of the lung cancer patient[10]

Exploratory Data Analysis (EDA):

EDA is an important initial step in ML to explore unfamiliar datasets with a sequence of analysis operations executed before applying ML to the dataset. We started with a sanity check by creating histograms of all features, as shown in Figure 2. Since we could not observe any outliers specifying this individual feature, we decided to use all data ranges rather than selecting specific conditions. As an alternative, we also looked over the correlation of the features by drawing pair plots based on matplotlib.[12] However, no clear correlations were observed. As the final step, we applied Peason's correlation which is normalized and a part of the feature selection procedure. This is presented as a heat map, in Figure 3. Pearson's correlation coefficients are the test statistics that measure the correlations of two continuous variables. If two features are highly correlated, we select only one feature for the model. In addition, visualization comparison is an essential process step for EDA.

We tried to look into every aspect of the dataset and visualize as many as possible. As a result, we could fully understand the basics of the statistics. Even so, since highly correlated features were not observed, we finally decided to include all features for ML analysis.

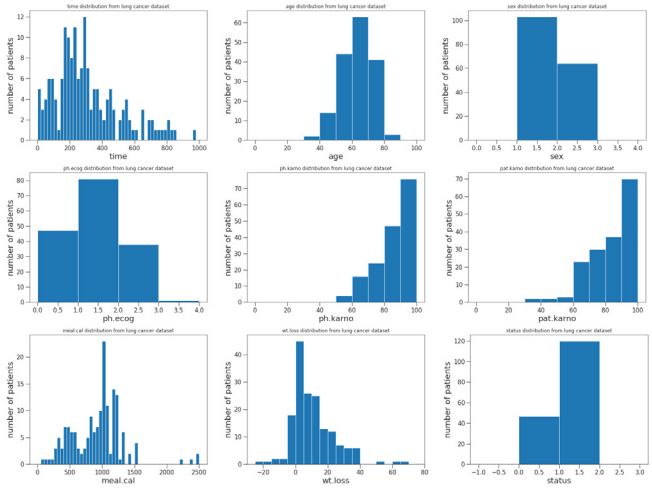

Figure 2: Histograms of the features; time, age, sex, ph.ecog, ph.krno, pat.karno, meal.cal, wt.loss, and status, from top left to bottom right

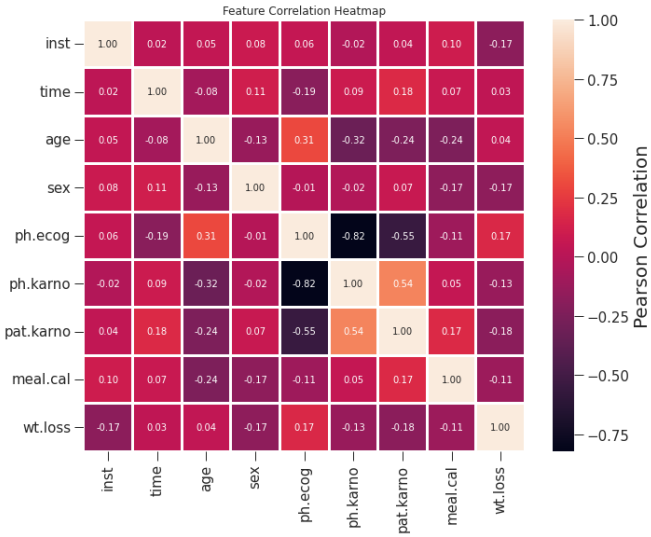

Figure 3: A heatmap of Pearson's correlation for the features from the dataset. The scale bar indicates the degree of correlation; 1 (bright orange) describes "fully correlation" and -1 (black) describes "no correlation"

Results and Discussion

Prediction with Machine Learning:

Machine Learning (ML) is broadly defined as the capability of a machine to imitate intelligent human behavior.[12,13] It is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without being explicitly programmed.[14] The objective of ML is to predict unseen data, learning from a large volume or multi-dimension of historical data. ML algorithms are used in a wide variety of areas in which conventional algorithms could be more efficiently applicable. The algorithms are based on mathematical equations that are implemented in codes.

DOI: 10.36838/v5i5.3

To process ML algorithms, we applied codes in the platform of Colab,[16] powered by python, a Google version of Jupyter notebook. We employed the notebook developed for Diabetes prediction via ML classification,[17] but we extensively modified its original body for our purpose. The Colab notebook used in this study is shared through the open link.[19] The ML algorithms applied are supervised learning methods; K-Nearest Neighbors (KNN), Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine (SVM), and Neural Networks, provided by sklearn as a python package.[18] The supervised learning method is an ML function that draws output models for prediction or performs prediction with classification by dividing inputs into train and test datasets. All methods were executed and compared according to the prediction performance. Details on each ML method can be found in the references.[1,13-15]

"status" is the target feature in this study, which is used to be either 1 or 2 as a value. Since typical target values in supervised learning are preferentially set up as either 0 or 1, we transformed the status features as 1 to 0 and 2 to 1. Then, we split the dataset into train and test samples. A typical ratio of the train-to-test is 7 to 3, but we chose 7.5 to 2.5 which is the default of the "train_test_split" function in Sklearn.[18] Since using stratify-parameter is important in "train_test_split" we keep the same distribution as a target in train and test datasets. For example, if variable y is a binary categorical value with 0 and 1, and there are 25% of zeros and 75% of ones, stratify will make sure that random split has 25% of 0's and 75% of 1's in each split sample.

We trained multiple ML algorithms mentioned above using the train samples. First, test samples were used to get some insights about accuracy and F1-score for the trained models. F1-score, also called F-score, is a machine learning evaluation metric that measures a model's accuracy. In addition, they were used to plot the confusion matrix. Then, we optimized hyperparameters using a grid search provided by Sklearn.[17] Once hyperparameters were optimized, we evaluated the performance of each algorithm using five-fold cross-validations. Some models provide a way to check feature importance, e.g., Random Forest, as shown in Figure 4. According to the feature importance plot, "time" and "wt. loss" are meaningful variables. In addition, Figure 5 shows the feature importance of Gradient Boosting, indicating "time" is the most meaningful variable, followed by "age".

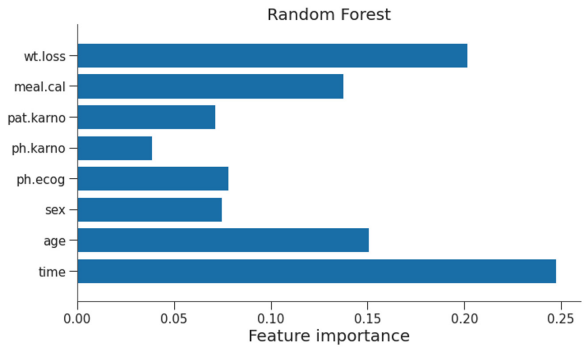

Figure 4: Feature importance from Random Forest

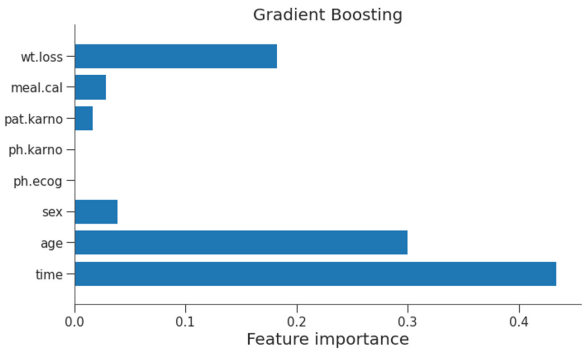

Figure 5: Feature importance from Gradient Boosting

In the same way, all models were finally evaluated by five-fold cross-validation. Figure 6 compares the obtained results. The performances of each algorithm are described based on the obtained accuracies and F1 scores. Table 1 summarizes the performance statistics. Ensemble models (e.g., Random Forest and Gradient Booting) typically perform better than non-ensemble models because ensemble learning is used to improve the performance of a model or reduce the likelihood of an unfortunate selection of a poor one. Similarly, our results show that Random Forest and SVM exhibit the best performance from the accuracy, while KNN performs best on F1-score. Other models indicate close to 70% accuracy but have lower (60%) F-scores overall.

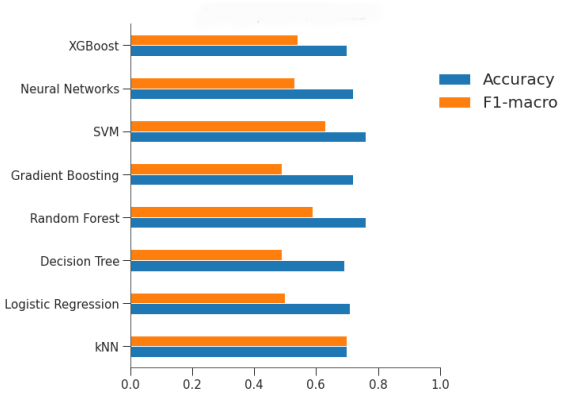

Figure 6: Model Performance based on accuracy and F1-score evaluated through the ML algorithms

Given the growing trend of machine learning (ML) application to the healthcare system, our attempt at the prediction performance for the NCCTG dataset using ML algorithms is worthwhile in addressing the advantages of ML technical applications. Our main goal is to understand the accessibility of ML-powered technology to lung cancer research as a part of oncology and to demonstrate its practical effect on decision-making for prognosis at the stage of cancer development. Unfortunately, the prediction accuracies from our evaluations are insignificant across all models. Indeed, we anticipated definite ML models to exhibit outstanding accuracies at the level that medical practitioners could satisfy, likely above 80%.[20] In comparison, the recent studies on breast cancer[21,22] and heart disease[23] provided ~ 92% and 86% of accuracies, respectively. At this moment, we believe that the lack of data statistics mainly contributed to the unsatisfied ou-

tcomes. In general, the small amount of dataset results in misclassifications and produces unstable and biased models. Although NCCTG data has a well-investigated structure and is organized the medical adequately, the dataset is likely not a fully informed structure for machine learning analysis. Its data does not contain any details on patients' generic measures, clinical notes, and reports as well.[24]

Even with this limitation, we explored the NCCTG dataset by employing three different processes of exploratory data analysis (EDA). We compared the correlations of the features for a sequence of ML analyses. After splitting the dataset into training and testing at the ratio of 7.5 and 2.5, the ML models were executed. Random forest, Decision Tree, and Gradient Boosting indicate that "time" is the most important key parameter. Similarly, XGBoost describes that "time" and "ph.ecog" are significant in the prediction. Using K (5) - fold cross-validation, the accuracy and F1-score of all models are compared as described in Table 1. Random Forest and KNN models show the best values. From this, we can conclude that "time" (the period of the observation) is the meaningful feature that can determine the survival rate of the advanced lung cancer patient. However, this outcome likely contradicts our initial goal because our analysis aimed to find a key feature for survival, depending on either the patient's or physician's assessments. The result directly tells of a "longer survival time," meaning "longer observation." It is irrelevant to the index examined with pat,krno, ph.karno, and ph.ecog". However, we speculate that this can be interpreted as the survival period supported by health practitioners' optimal clinical treatment. This hypothesis is likely feasible according to the recent study24 that shows the correlation of the survival rate of lung cancer with the time of the patient's sustenance.

Table 1: The Detailed Statistics of Performance for each algorithm

Algorithms	Accuracy	Std deviation	F1-score	Std deviation
KNN	0.70	+/- 0.04	0.70	+/- 0.08
Logistic Regression	0.71	+/- 0.02	0.50	+/- 0.08
Decision Tree	0.69	+/- 0.01	0.49	+/- 0.08
Random Forest	0.76	+/- 0.06	0.59	+/- 0.12
Gradient Boosting	0.72	+/- 0.03	0.49	+/- 0.07
SVM	0.76	+/- 0.05	0.63	+/- 0.09
Neural Networks	0.72	+/- 0.01	0.53	+/- 0.07

Of course, if a relatively large volume of the well-structured dataset was provided, we could achieve high accuracy in providing the specific predictive outcomes. Furthermore, a large dataset allows sufficient partitioning into training and testing sets, thus leading to reasonable validation through an extended process of K-fold cross-validation. Although our study has several limitations, we hope this initial attempt presents the potential to improve prediction performance for lung cancer research employing ML algorithms. In addition, this study is another effort to specify machine learning as an enabling tool to prevent cancer deaths effectively. The recent study[24] clearly shows that ML powered screening process provides better performance, compared to PLCOm2012 (2012 Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial Model), a lung cancer risk prediction model validated by long-term clinical research.

Conclusion

By applying multiple machine learning algorithms to the lung cancer dataset from NCCTG, we attempt to predict important features that play a pivotal role in determining cancer survival. It also demonstrates the practical effect of ML on decision-making for cancer prognosis. Our studies show that overall accuracy of ~ 75% is acquired, and "time" is the most decisive parameter. "Time" can be interpreted as "proper clinical procedure." We suspect that our dataset lacks medical statistics, resulting in not-high enough accuracy. To obtain reliable results regarding the predicted performance of ML models, a large volume of the clinical dataset is required. We also propose that an understanding of how sophisticated a dataset is needed for higher accuracy (at least above 80%) should be advanced. Finally, how much the amount of a dataset is required to optimize the prediction performance should be answered. These proposed studies will be carried out shortly.

Acknowledgments

I am thankful to my mentor for guiding me on this research project. I have no disclosures to make.

References

1. Shailaja, K.; Seetharamulu, B.; Jabbar M. A.; "Machine Learning in Healthcare: A Review," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, **2018**, pp. 910-914, doi: 10.1109/ICECA.2018.8474918.
2. Qayyum, A.; Qadir, J.; Bilal M.; Al-Fuqaha, A.; "Secure and Robust Machine Learning for Healthcare: A Survey", in IEEE Reviews in Biomedical Engineering, **2021** vol. 14, pp. 156-180, doi: 10.1109/RBME.2020.3013489.
3. Andrey Koptelov; "*Machine learning in healthcare: a complete overvi ew*," https://www.itransition.com/machine-learning/healthcare (2022)
4. Kourou, K.; Exarchos, K. P.; Papaloukas, C.; Sakaloglou, P.; Exarchos, T.; Fotiadis, D. I.; "Applied machined learning in cancer research: A systematic review for patient diagnosis, classification and prognosis," Computational and Structural Biotechnology Journal 19, **2021**, 5546–5555
5. Cruz, J. A.; Wishart, DS.; "Applications of machine learning in cancer prediction and prognosis," Cancer Informat 2006;2:59.
6. American Cancer Society (https://www.cancer.org/), NIH (National Institutes of Health), National Cancer Institute (https://www.cancer.gov/)
7. Loprinzi, C. L.; Laurie, J. A.; Wieand, H. S.; Krook, J. E.; Novotny. P. J.; Kugler. J. W.; Bartel, J.; Law, M.; Bateman, M.; Klatt,N. E.; "Prospective evaluation of prognostic variables from patient-completed questionnaires," North Central Cancer Treatment Group. J Clin Oncol. **1994** Mar;12(3):601-7. doi: 10.1200/ JCO.1994.12.3.601. PMID: 8120560.
8. *World Cancer Research Fund International, Worldwide cancer data:* https://www.wcrf.org/cancer-trends/worldwide-cancer-data/
9. *Center for Disease Control and Prevention: Basic information about lung cancer*: https://www.cdc.gov/cancer/lung/basic_info/index.htm
10. *American Society of Clinical Oncology: Cancer.NET, Lung Cancer-*

DOI: 10.36838/v5i5.3

Non-Small Cell: Statistics: https://www.cancer.net/can cer-t% C3 %BDpes/lung-cancer-non-small-cell
11. *Daxplore demos and Kaggle* https://dataset.lixoft.com/data-set-examples/ncctg-lung-cancer-data-set/, https://www.kaggle.com/datasets/ukveteran/ncctg-lung-cancer-data
12. *Matplotlib: visualization with python*: https://matplotlib.org/
13. https://en.wikipedia.org/wiki/Machine_learning
14. Mahesh, B.; "Machine Learning Algorithms—A Review," International Journal of Science and Research, **2020**, 9, 381-386.
15. Ray, S.; "A Quick Review of Machine Learning Algorithms," 2019 *International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, Faridabad, India, 2019, pp. 9, pp. 35-39, doi: 10.1109/COMITCon.2019.8862451.
16. *Google Colaboratory*: https://colab.research.google.com/
17. *https://github.com/Vikasdubey0551/Diabetes-classification-Machine-learning*
18. *Sklearn (scikit-learn)*: https://scikit-learn.org/stable/
19. *The google colaboratory notebook applied in this study is shared via this link*: https://colab.research.google.com/drive/1ZqRB4nkWqhWqhQfyCQnAlipW69_ohgg7OLg?usp=sharing
20. Iyer, A.; Jeyalatha, S.; Sumbaly, R.; "Diagnosis of Diabetes Using Classification Mining Techniques," International Journal of Data Mining & Knowledge Management Process (IJDKP), **2015** vol.5, pp. 1-14.
21. Williams, K.; Idowu, P. A.; Balogun, J. A.; Oluwaranti, A.; "Breast cancer risk prediction using data mining classification techniques," **2015** Transactions on Networks and Communications, vol.2, pp.1-11.
22. Senturk, Z. K.; Kara, R.; "Breast cancer Diagnosis via Data Mining: Performance Analysis of Seven Different Algorithms," **2014** Computer Science & Engineering", vol.1, pp.1-10.
23. Vembandasamy, K.; Sasipriya, R.; Deepa, E.; "Heart Diseases Detection Using Naive Bayes Algorithm," **2015** vol.2, pp. 441-444.
24. Gould, M. K.; Huang, B. Z.; Tammemagi, M. C.; Kinar, Y.; Shiff, R.; "Machine Learning for Early Lung Cancer Identification Using Routine Clinical and Laboratory Data," **2021** AM J Respir Crit Care Med Vol 204, Iss 4, pp 445-453, Aug 15.

Authors

Patrick Junhee Kim is a junior at Asia Pacific International School, interested in chemical engineering and biology. He hopes to build his career as a pioneer in the bioengineering field using state-of-the-art IT technology like machine learning and artificial intelligence.

Andrew Seungwoo Youn is a junior at Asia Pacific International School eager to join the chemical and nanotechnical industry. He wants to run a venture company in bioinformatics or chemistry. He hopes to major in chemistry and biology in college.